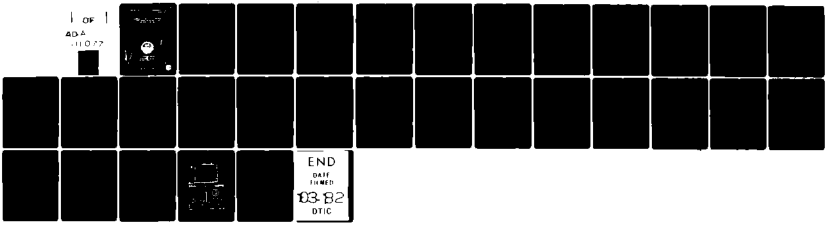
AD-A111 077 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
PERFORMANCE TEST OBJECTIVITY: A COMPARISON OF RATER ACCURACY AN--ETC(U)
FEB 82 W A NUGENT, G J LAABS, R C PANELL
UNCLASSIFIED NPRDC-TR-82-30 NL

1 OF 1
AD A
111077



END
DATE
FILMED
03-82
DTIC

NPRDC TR 82-30

FEBRUARY 1982

PERFORMANCE TEST OBJECTIVITY: A COMPARISON OF RATER ACCURACY AND RELIABILITY USING THREE OBSERVATION FORMS

NAVY PERSONNEL RESEARCH
AND
DEVELOPMENT CENTER
San Diego, California 92152

NPRDC TR 82-30 February 1982

PERFORMANCE TEST OBJECTIVITY: A COMPARISON OF RATER ACCURACY AND RELIABILITY USING THREE OBSERVATION FORMS

William A. Nugent
Gerald J. Laabs
Robert C. Panell

Reviewed by
E. G. Aiken

Released by
James F. Kelly, Jr.
Commanding Officer

Navy Personnel Research and Development Center
San Diego, California 92152

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
NPRDC TR 82-30

2. GOVT ACCESSION NO.
AD-A111 077

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
PERFORMANCE TEST OBJECTIVITY: A COMPARISON OF RATER ACCURACY AND RELIABILITY USING THREE OBSERVATION FORMS

5. TYPE OF REPORT & PERIOD COVERED
Technical Report
Dec 1979--Apr 1980

6. PERFORMING ORG. REPORT NUMBER
15-81-22

7. AUTHOR(s)
William A. Nugent
Gerald J. Laabs
Robert C. Panell

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Navy Personnel Research and Development Center
San Diego, California 92152

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
ZF63-522-002-03.40,
ZF63-522-001-014-03.01

11. CONTROLLING OFFICE NAME AND ADDRESS
Navy Personnel Research and Development Center
San Diego, California 92152

12. REPORT DATE
February 1982

13. NUMBER OF PAGES
30

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Job performance evaluation techniques	Job proficiency measurement devices
Rater guidance materials	Videotaped evaluation problems

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The study examined two variables that may influence the consistency and accuracy of rater's judgments in evaluating job performance: (1) the precision with which behaviors to be observed and evaluated are specified on a performance observation form and (2) the level of proficiency of the rater at the task being evaluated.

Videotapes were prepared that depicted both passing and failing performances in the use of two types of electronic test equipments. The videotapes were observed in two experiments by raters who used a structured, semistructured, or unstructured performance observation form. Rater skill level was determined by the score the rater obtained on a performance test that consisted of the same types of electrical measurement problems as shown on the videotape. For both experiments, the presence of at least some structure in a performance observation form produced more accurate and reliable evaluations of job task performance than did a form with no structure. Within the range of rater skill tested, results showed that the level of skill proficiency that raters have with a particular type of electronic test equipment is largely independent of their ability to judge accurately and consistently the performance of others in using the same equipment.

FOREWORD

This research and development was initiated under subproject ZF63-522-002-03.40 (Techniques for the Measurement of Job Performance) and completed under subproject ZF63-522-001-014-03.01 (Component Analysis of Resources and Readiness Systems). The work, which was sponsored by the Chief of Naval Education and Training, was initiated by the Navy Personnel Research and Development Center to develop techniques and instruments for the evaluation of job performance capabilities of Navy personnel.

The graphic symbolic simulation version of a hands-on oscilloscope proficiency test described in this report was developed under contract N66001-79-C409 by Kinton, Inc., Alexandria, Virginia.

Appreciation is expressed to the staff of the Fleet Anti-submarine Warfare Training Center, Pacific for providing access to students and classroom facilities. Special acknowledgement is made to Mr. Robert C. Panell for his work in planning this research, which he was not able to see to its completion. Prior to his untimely death, Mr. Panell was a member of the Performance Measurement Group at the Navy Personnel Research and Development Center.

Results of this research and development are intended for use by organizations responsible for assessing job performance in general; and for the Personnel and Training Evaluation Program, Naval Technical Training Command, and Naval Education and Training Command, in particular.

JAMES F. KELLY, Jr.
Commanding Officer

JAMES J. REGAN
Technical Director

SUMMARY

Problem

Because hands-on performance tests tend to have high face validity, it is typically assumed that they are, by their very nature, valid and reliable. Therefore, judgment accuracy (i.e., the degree to which the rater's judgments reflect how individuals actually performed) and judgment reliability (i.e., the degree to which raters agree when they observe the same set of individuals) are rarely investigated. Two variables that may influence the accuracy and consistency of rater judgments are (1) the degree of structure in the test observation form (i.e., the precision with which the behaviors to be observed are specified) and (2) the level of proficiency of the rater at the task being evaluated.

Objectives

The objectives of the present effort were to (1) determine the extent to which the degree of structure provided by test observation forms influences rater judgments and (2) examine the relationship between a rater's accuracy and skill level in evaluating a hands-on job performance.

Method

Videotapes were prepared that depicted varying levels of performance in the use of two types of electronic test equipments--a volt-ohm-milliammeter (VOM) and an oscilloscope. These videotapes were observed in two experiments by raters who used either a structured, semistructured, or unstructured form. In these experiments, raters who had been classified as either high- or low-skill proficient on VOM and oscilloscope proficiency tests observed videotapes depicting passing or failing performances on the use of the VOM and oscilloscope.

Results

Judgment accuracy was determined by the percent of rater agreement with a predetermined pass/fail criterion for each videotape performance sequence. In an analysis of variance of these data, a significant effect for observation form was found in both experiments. In Experiment I, raters' judgments on the structured observation form were significantly more accurate than were those obtained on the semistructured or unstructured forms. In Experiment II, raters' judgments on both the structured and semistructured forms were more accurate than were those obtained on the unstructured form. Rater proficiency did not affect the accuracy of judgments.

Judgment reliability was estimated by the amount of interrater agreement. In Experiment I, it was found that raters who used the structured form showed significantly higher reliability (r = .90) when compared to the semistructured (r = .58) and unstructured (r = .30) forms. Reliability coefficients for Experiment II did not differ significantly, although the structured and semistructured forms showed higher reliability (r = .67 and .72 respectively) than did the unstructured form (r = .32). No differences were found in reliability as a function of rater skill proficiency.

Conclusions

1. The anticipated results that raters who were more highly skilled in the operation of a particular type of electronic test equipment would be more accurate and consistent in evaluating the performance of others in using the same equipment did not materialize.

This means that having above average skill in a given task area does not automatically guarantee superior performance in terms of either rater accuracy or reliability.

2. Evaluations of job task performance are more accurate and reliable when made by using a performance observation form with at least some structure than one with no structure.

3. The lessened effect of the structured observation form on the more complex oscilloscope task indicates that a trade-off may exist between the information-processing demands on the observer and the use of a highly detailed form to evaluate job task performance. That is, the advantage in using a detailed form to observe a task may be offset either as the complexity of the task or number of interchangeable steps in the task increases.

Recommendations

1. In developing forms to observe and evaluate job performance, some structure (i.e., detailed listing of procedural steps) should be provided.

2. The amount of structure provided in a performance observation form should be determined on the basis of the complexity of the task to be observed. That is, a less-structured form may be acceptable as the complexity of the task or number of interchangeable steps in the task increases. Further research should be conducted to determine the exact nature of the trade-off between complexity and performance observation form specificity.

3. Use of the structured observation technique is not restricted to the evaluation of actual on-the-job performance. The use of this technique should be tested to determine its suitability for evaluating trainee performance in formal Navy training schools.

CONTENTS

INTRODUCTION

Problem

The Navy uses a variety of techniques for evaluating job performance skills. When a given skill is evaluated on the basis of a hands-on performance test, for example, it is typically assumed that, if the test seems relevant to the behavior domain to be measured (i.e., has high face validity), the evaluative procedures and resultant data associated with the test are both valid and reliable. If, however, ambiguity exists in the performance steps to be observed and evaluated, inaccurate observations or low agreement among observers or raters may result. In addition, a lack of objectivity (i.e., a rater's dependence on subjective judgment) may adversely affect the validity and reliability of the evaluation procedure. Two factors related to objectivity in the evaluation of job performance were investigated in the present study: (l) the degree of structure in the test observation form (i.e., the precision with which the behaviors to be observed are specified) and (2) the level of proficiency of the rater at the task being evaluated.

Background

Although techniques are available for optimizing the level of objectivity when evaluating job performance (written multiple-choice tests and computer-administered performance tests with automated scoring routines are two examples), there are many instances when the use of such techniques are not feasible or practical. As a result, job performance evaluations in the Navy are typically conducted by senior supervisory personnel or job incumbents who have demonstrated a satisfactory level of proficiency in the task to be evaluated.

Given this state of affairs, the current effort defined objectivity as the degree to which raters base their evaluations on external standards rather than personal judgment. In a controlled laboratory situation, performance test objectivity can be easily evaluated in terms of the accuracy of the observations in comparison to a predetermined criteria. When such criteria are unavailable or ill-defined, as in the case of observing performance on the job, objectivity can be measured in the following manner (Pickering & Anderson, 1976):

> A test is objective when different examiners can use it to observe the same individuals at the same time and obtain comparable results. Another indication of objectivity would be when the same examiner can use the test to make the same, or nearly the same, observations when presented with identical situations at different points in time. (p. 11)

In an early investigation on the objectivity of raters judgments, Siegel (1954) examined the consistency of the results obtained when a group of trained observers evaluated the same filmed performance with a 1-month interval between showings. It was found that the percent agreement of individual rater's judgments between the two evaluation periods (i.e., intrarater consistency) varied from 64 to 100 percent. Siegel concluded that intrarater consistency should be determined before raters are assigned to evaluate performance on a job sample test. Furthermore, Siegel concluded that, if all raters show low consistency in judgments, then either the performance test itself is inadequate or rater training, in terms of the task to be evaluated, has been poor.

More recent job performance tests developed for the U.S. Army (McClusky, Trepagnier, Cleary, & Tripp, 1975) covered a variety of tasks, including preparing and firing weapons and installing and recovering mines. Four raters independently evaluated 15 recruits as the tasks were performed. On all of the performance tests, rater agreement was found to be very low. The following factors contributed to the low rater agreement:

1. Some test measures consisted of ambiguous statements that were open to interpretation and subjective judgment.

2. In some instances, two actions were included in one performance step. This resulted in confusion as to whether completion of one or both steps was required for a "yes" answer.

3. When a particular sequence of steps was required, it was not always clear whether steps performed out of sequence should be scored as correct or incorrect.

These studies suggest two areas related to the objectivity of hands-on performance tests that have not received very much attention: (1) the amount of structure provided in a performance observation form and (2) the skill proficiency of the rater. Without specific guidelines on what steps or processes to observe, a rater is forced to make subjective judgments that are based on internal standards. Raters should not be expected to evaluate steps they cannot see, such as those involved in evaluating a mental process, and each step should be clearly stated. When several processes are observed and evaluated as a single step or there is ambiguity as to what constitutes a performance step, it becomes more difficult to obtain consistent judgments across raters. As a consequence, the current effort hypothesized that the more structure included in an observation form (i.e., the more precision with which performance steps are specified), the more the raters should agree on completion of steps in a problem. Another important variable that may interact with test objectivity is the expertise of the rater. The degree of experience that a rater has with a particular system or equipment may influence his judgment of how others use it. In the Navy's Personnel Qualifications Standards (PQS) program, for example, job performance evaluations are conducted by senior supervisory personnel or by job incumbents who have, presumably, demonstrated proficiency in the section to be evaluated. It is assumed that, when raters are qualified in this manner, they do not require structured observation forms to aid them in measuring specific skills.

Objectives

The objectives of this effort were to (1) determine the extent to which the degree of structure provided by test observation forms influences rater judgments and (2) examine the relationship between the rater's ability to evaluate accurately the performance of others and the rater's proficiency level. Two experiments were conducted in this effort. The first was an examination of the effects of performance evaluation format and rater skill proficiency on the accuracy and reliability of performance judgments using a relatively simple test equipment operator task; the second was a replication of the first, but using a more complex test equipment operator task.

GENERAL APPROACH

Stimulus Materials

To determine the accuracy and reliability of the judgments of raters, the behavior they are required to observe and evaluate must be held constant. One way this can be accomplished is by videotaping the test behavior so that any variation in evaluations is due to rater or observation form differences and not to test performance differences. To study job performance evaluations for the present effort, videotape scripts were prepared to depict both passing and failing performances on two standard Navy test equipments--a Simpson Model 260 volt-ohm-milliammeter (VOM) and the AN/USM-281A oscilloscope. The videotape scripting procedure ensured that the performances had predetermined outcomes (i.e., problems were either passed or failed), thereby providing criteria against which the accuracy of rater's evaluations can be compared.

Personnel from Navy Personnel Research and Development Center (NAVPERSRANDCEN) served as the test equipment operators (i.e., examinees) in the production of the videotaped performances. In each performance, the examinee was told what electrical measurement to perform, performed the steps needed to solve the measurement problem, announced that the measurement had been completed, and gave a report of the final readings obtained.

Performance Observation Forms

In each experiment, the participants used one of three different observation forms--a structured, semistructured, or unstructured form--to evaluate the videotaped performances of the electrical measurements.

The unstructured form, which was modeled after a part of the Navy's Personnel Qualifications Standards (PQS) program, required the rater to make a pass or fail judgment based on overall task performance for each videotaped problem and contained spaces for recording any errors detected. No step-by-step procedures were provided to guide the evaluations nor were any criteria specified to define passing or failing performance. Thus, a rater's subjective judgment could play a large part in making the overall evaluation.

The semistructured form was similar to performance observation forms developed and used in the past (e.g., Laabs, Panell, & Pickering, 1977). The specific form used in this effort was adapted from those used in a self-paced electronic test equipment course that is currently administered at the Submarine Training Center, Charleston, South Carolina. This form required the rater to evaluate performance against a set of one or more procedural steps organized under four structured areas of performance (i.e., preliminary adjustments, control settings, waveform analyses, and safety). Each area contained both critical and incidental criteria for determining passing or failing performance. Maximum numerical values were provided for each area and raters were instructed to award points up to the maximum specified using their own judgments. When the performance task was completed, the rater summed the individual point values assigned to determine whether passing criteria (i.e., 75% correct) had been met. Specifying the performance areas to which some portion of a predetermined point value was to be assigned removed some of the subjectivity involved when performance was observed and evaluated using an unstructured form.

The structured rating form, which was developed specifically for the present study, required each measurement problem to be analyzed in a series of small, discrete actions called checkpoints. Each checkpoint was expressed as a statement of the correct action to be performed. The checkpoints were arranged sequentially to form a performance checklist with which an observer could evaluate performance. In addition to performance checkpoints, the structured observation form contained a graphic illustration of the test equipment face. This illustration was included so that the position of control settings, the location of lead connections, and actual readings obtained could be easily marked on the response form. Failure to perform the correct action at any checkpoint led to failure on the measurement problem. Thus, the structured observation form left relatively little room for subjective judgments.

Examples of the three types of performance observation forms used in Experiments I and II are provided in Appendices A and B respectively.

Rater Skill Level

Rater skill level was determined by the score the rater obtained on a performance test that consisted of the same types of electrical measurement problems shown on the videotape. For Experiment I, the rater performed four hands-on problems using the VOM. A member of the research staff used the structured observation form to evaluate rater skill level on these hands-on problems. For Experiment II, participants were given a graphic symbolic simulation version of a hands-on oscilloscope proficiency test. This test, which included 21 problems, preserved the stimulus aspects of an actual hands-on proficiency test through the use of graphics. Responses were made either by circling appropriate control settings or by calculating values associated with simulated waveforms. Previous research involving the symbolic simulation test (Laabs, Nugent, & Bearden, 1981) indicated that it was comparable to a hands-on oscilloscope test in terms of discriminant validity and classification consistency and was superior to the hands-on version with respect to overall test reliability.

In each experiment, two skill proficiency categories were established for the rater expertise variable that roughly divided each sample into two groups. For Experiment I, raters passing two or more of the four problems were assigned to the high-skill proficient group; and the remainder, to the low-skill proficient group. For Experiment II, raters were assigned to the high-skill proficient group if they responded correctly to 12 or more of the 21 problems. In both cases, this classification procedure resulted in a rater grouping that differed significantly in terms of skill proficiency level (i.e., $t = 16.7$, $df = 76$, $p < .001$, and $t = 12.0$, $df = 70$, $p < .001$ for Experiments I and II respectively).

EXPERIMENT I: EVALUATING PERFORMANCE ON THE VOLT-OHM-MILLIAMMETER

Test Problems

In the videotaped performances in Experiment I, examinees made four types of electrical measurements (i.e., negative DC voltage, positive DC voltage, resistance across a signal generator, and resistance across a resistor). Three videotaped performances were prepared for each type of electrical measurement--one in which the entire measurement was performed correctly and two in which various procedural errors were committed. Thus, stimulus materials consisted of 12 videotaped performances, four of which were correct performances; and eight, incorrect. For presentation purposes, the videotaped performances were combined on a master videotape in which the problem sequence (negative voltage, positive voltage, signal generator resistance, resistor resistance) was repeated three times. The correct and incorrect performances associated with each type of electrical measurement occurred at random within each sequence.

Problems were performed on a Simpson Model 260 VOM and, for those requiring electrical outputs for measurement, a Hydrotronics test signal generator.

Procedures

Half the rater sample received the VOM proficiency test prior to viewing the videotaped presentations; and half, after viewing them. This was done to allow the assessment of possible effects arising from completing the VOM proficiency test prior to evaluating the videotaped performances. In both conditions, raters were tested individually on the skill proficiency test and each rater was assigned one of the three performance observation forms on a random basis.

Testing was conducted in an experimental laboratory at NAVPERSRANDCEN. Raters conducted their evaluations of the videotaped performances in groups of two or three at individual television monitor carrels. Prior to evaluating the performances, raters were given a practice session to enable them to become familiar with the videotape format as well as the performance evaluation form they had been assigned. The raters viewed each videotape performance, consisting of a single electrical measurement problem, only once. Following each performance, raters were given 30 seconds to complete their evaluation forms.

Sample

A total of 15 instructors and 63 students from the Fleet Anti-Submarine Warfare Training Center, Pacific (FLEASWTRACENPAC) participated in the study. The students were either designated sonar technicians (STs) or were undergoing initial "A" school training in that rating.

Of the 78 raters tested, 28, 26, and 24 raters were assigned on a random basis to the structured, semistructured, and unstructured performance observation forms, respectively. In the absence of a well-defined performance standard on the VOM proficiency test, the median test score was used as the criterion for classifying raters on the skill proficiency variable. Using this criterion, 16 of the raters who used the structured form were classified as high skill proficient and 12 as low skill proficient, compared to 14 and 12 for the semistructured form, and 12 and 12 for the unstructured form.

Results

In six of the eight videotaped problems considered as failing performances for Experiment I, examinees made major procedural errors that resulted in incorrect meter readings (i.e., erroneous solutions). On the two remaining problems, the meter readings obtained by examinees were in the correct response ranges, but they had committed minor procedural errors that affected the accuracy of these measurements (e.g., meter not set to zero using calibration setscrew). Because not enough points could be deducted for those types of errors on the semistructured observation form to warrant a failing score, the predetermined criterion for these performances could not be applied across all three forms. Therefore, responses to to these two failing videotape performances were excluded from further analyses; experimental data were generated from ten videotaped measurement problems--six failing performances and four passing performances.

Accuracy

The percent of rater agreement with the predetermined pass/fail criterion for the ten videotaped performances was calculated, and a form-by-skill-by-order analysis of variance (ANOVA) was performed. The main effect for performance observation form was found to be statistically significant (F(2,66) = 25.92, p < .001). A Scheffe (1953) post-hoc analysis revealed that agreement when using the structured form (97.1%) differed significantly from that when using the semistructured (80.8%) and unstructured (76.7%) forms (p < .01). The effects associated with rater skill proficiency level and test presentation order, as well as all interactions, failed to reach statistical significance. When an estimate of the overall strength of association was calculated (Hays, 1973, p. 512), it was found that 39 percent of the variance in the criterion agreement variable was accounted for by the performance observation form used.

The results of the same type of analysis conducted separately for the four passing and six failing videotaped performances support those obtained for the overall analysis in that the effect for the observation forms was the only statistically significant finding. For the passing videotaped performances, it was found that criterion agreement when using either the structured or semistructured forms (95.5 and 88.5% respectively) differed significantly (p < .01) from that when using the unstructured form (69.8%). The criterion agreement on the failing performances was found to be significantly higher (p < .01) when the structured form was used (98.2%) than when the semistructured or unstructured forms were used (75.6 and 81.2% respectively).

Observation Errors on Failing Performances

These findings indicate that product judgments (i.e., assigning pass/fail scores) are best made using the structured performance observation form. These data do not, however, fully describe the state of affairs in using the different observation forms, because they do not reflect the errors made in observing the process or procedural steps in the electrical measurement problems. For example, assignment of a failing score that was in agreement with the predetermined criterion could be made for the wrong reason. This might involve a missed-event (failure to identify an incorrectly performed procedural step), coupled with a false alarm (identifying a correctly performed procedural step as incorrect). Although the three performance observation forms were not designed to provide equivalent amounts of information for the process judgments, it was felt that a more detailed examination of the errors made observing the six videotapes of incorrect performances would be useful.

Table 1 shows the average percent of missed-event errors for the three observation forms. For the structured form, this meant that an incorrect step was marked as correct; for the semistructured form, that points were not deducted for an incorrect step; and for the unstructured form, that an error was not written down. Since there was no way of determining whether the rater observed the incorrectly performed step and neglected to enter it on the observation form, the missed-event error rate for the unstructured form might be inflated. Nevertheless, a much lower percentage of missed-event errors was associated with the structured form, which supports the findings on rater criterion agreement across the performance observation forms.

Table 1

Mean Percent of Missed-event Errors in Failing Performances on the Volt-ohm-milliammeter by Observation Form

	Rating Form		
Item	Structured	Semistructured	Unstructured
M	7.1	20.2	50.5
SD	8.6	13.3	28.2

Comparable analysis of the false-alarm error rate for the three performance observation forms was not possible because these data could not be systematically identified from the semistructured form.

Reliability

An estimate of interrater reliability was calculated for each performance observation form through application of the dichotomous pass/fail responses to an ANOVA technique that yields an intraclass correlation (Shrout & Fleiss, 1979). It was found that the structured form showed the highest reliability ($r = .90$). The reliability coefficients for the semistructured and unstructured forms were .58 and .30 respectively. These coefficients are estimates of the reliability of a single rater. The differences in reliability coefficients were tested by a Chi-square analysis (Snedecor & Cochran, 1980, p. 187) and found to be statistically significant ($\chi^2 = 16.96$, df = 2, $p < .001$). A post-hoc analysis of these reliability coefficient values revealed that the structured rating form differed significantly from the semistructured and unstructured forms ($p < .01$).

No significant differences were found in interrater reliability as a function of rater skill proficiency.

EXPERIMENT II: EVALUATING PERFORMANCE ON THE OSCILLOSCOPE

Test Problems

The videotaped performances in Experiment II consisted of 16 measurement problems in the areas of amplitude, frequency, pulse duration, superimposed DC voltage, and probe calibration. There were four problems in each of the first three areas and two problems in each of the latter two areas. Half of the oscilloscope problems were performed correctly; and half were not. Two of the 16 videotaped performances (i.e., measuring amplitude and frequency) were used as practice problems. The remaining 14 performances were presented to the rater sample in a random order.

All videotaped problems were performed on a standard Navy dual-trace oscilloscope, Model AN/USM-281A. In addition, for the problems involving the amplitude, frequency, pulse duration, and superimposed DC voltage, a Continental Specialties Corporation Model 2001 test signal generator was used to provide the waveforms for the electrical measurements.

Procedures

As in Experiment I, half the rater sample received the oscilloscope symbolic simulation test prior to viewing the videotaped performances; and the other half, after viewing the performances. In both conditions, the symbolic simulation test was administered to raters in a group se*ting and each rater was assigned to one of three performance observation forms on a random basis.

Testing was conducted in a student classroom at FLEASWTRACENPAC. Raters evaluated the videotaped performances in groups of two or three at individual television monitor carrels. Prior to evaluating the 14 videotaped performances, raters were given a practice session that consisted of two parts. First, to familiarize all raters with the test equipment used in the experiment, a videotape was shown that provided instruction on the location and interpretation of various oscilloscope switches and controls. Second, to familiarize them with the format of the videotaped performances and their assigned performance observation forms, raters viewed the two practice problems mentioned earlier.

The raters viewed each videotaped performance, consisting of a single measurement or probe calibration problem, only once. At the end of each performance, raters were shown a videotaped segment that reviewed the final position of all oscilloscope switch and control settings. This segment was followed by a 30-second blank period on the videotape to allow raters to complete their performance evaluation forms. The forms were collected when raters had completed their evaluation of the final videotaped performance.

Sample

A total of 8 instructors and 64 students from FLEASWTRACENPAC participated in the study. The students in the study were designatd STs, who were either currently enrolled in advanced sonar maintenance training courses or awaiting assignment to these courses.

Of the 72 raters tested, 24 raters were assigned on a random basis to each of the three performance evaluation forms. As was the case in the first experiment, a well-defined performance standard was not available to differentiate among rater skill proficiency levels based on the oscilloscope symbolic simulation test. Therefore, the median test score was used again as the criterion for classifying raters on the skill proficiency variable. Using this criterion, 36 raters were classified as high-skill proficient; and the remainder, as low-skill proficient.

Results

Of the 14 videotaped performances evaluated by the raters, the two superimposed DC voltage measurements were discarded because a critical performance step had been inadvertently omitted in the production of the videotape. Thus, experimental data were generated from 12 videotaped measurement problems.

Accuracy

The percent of rater agreement with the predetermined pass/fail criterion for the 12 videotaped performances was calculated and a form-by-skill-by-order ANOVA was performed. For the overall videotaped performances, the main effect for observation form was found to be statistically significant (F (2,60) = 32.19, p < .001). A Scheffe (1953) post-hoc analysis of the overall criterion agreement values for the three rating forms showed that rater agreement on the structured and semistructured forms (90.3 and 92.0% respectively) differed significantly (p < .01) from the unstructured form (70.1%). The effects of rater skill level and test presentation order, as well as all interactions, failed to reach statistical significance. An estimate of the overall strength of association between observation form and criterion agreement was also calculated. It was found that 46 percent of the variance in the criterion variable can be accounted for by the performance observation forms used.

Additional analyses of the same type were conducted separately for the six passing and six failing videotaped performances on the oscilloscope. The main effect for observation form on the failing performances was the only statistically significant finding (F (2,60) = 48.86, p < .001). A comparison of the mean criterion agreement values for the failing performances showed that rater agreement on the structured and semistructured forms (95.8 and 91.0% respectively) differed significantly (p < .01) from the unstructured form (59.0%).

Observation Errors on Failing Performances

Table 2 shows the average percent of missed-event errors for the three rating forms used in Experiment II. Results obtained from this analysis were similar to those found in the Experiment I. Raters who used the structured performance evaluation form had a lower missed-event error rate than did raters who used the semistructured and unstructured forms.

Table 2

Mean Percent of Missed-event Errors in Failing Performances on the Oscilloscope by Observation Form

	Rating Form		
Item	Structured	Semistructured	Unstructured
M	10.9	14.6	57.8
SD	13.9	14.1	21.8

Reliability

The estimates of interrater reliability calculated for the three performance evaluation forms used in Experiment II did not differ significantly, although the structured and semistructured forms had higher estimated reliability (r = .671 and .721 respectively) than did the unstructured form (r = .317). Also, results of a comparison of interrater reliability coefficients in terms of rater skill proficiency level within each observation form were not significant.

DISCUSSION

The results show that, within the range of rater skill tested, the level of skill proficiency that raters have with a particular type of test equipment is largely independent of their ability to accurately and consistently judge the performance of others in using the same equipment. This means that having above average skill in a given task area does not automatically guarantee superior performance in terms of rater accuracy or reliability.

The failure of the skill proficiency variable to emerge as a significant source of variance in both experiments does not, however, indicate that the skill proficiency of a rater should be ignored when evaluating job performance. Because a well-defined performance standard was not available to differentiate among rater skill levels on either the VOM or oscilloscope proficiency tests, an arbitrary criterion (i.e., median score) was used to classify rates as high or low skill proficient. Use of this classification procedure did produce groups that differed significantly in terms of mean performance on the skill proficiency test. However, an inspection of the high and low skill-group test score distributions for each experiment revealed that outlying scores maximized the group differences. Furthermore, it was found that 65 percent of the rater sample on the VOM proficiency test fell within 1 point above or 1 point below the skill classification criterion; and that 53 percent of the sample were within 3 points above or 2 points below the classification criterion on the oscilloscope proficiency test. Thus, it would appear that the rater skill groups were not clearly differentiated in terms of their level of proficiency near the classification criterion.

With respect to varying the amount of structure in the observation forms, results indicate that the presence of at least some structure produces more accurate as well as more reliable judgments than when there is no structure. For example, rater accuracy values obtained in the first experiment show a drop from almost perfect agreement with the overall pass/fail criterion when raters used the structured observation form (i.e., 97%) to about 77 percent when raters used the unstructured form. The listing of unambiguous step-by-step procedures on the structured form also resulted in significantly higher interrater reliability. With less structure in the observation forms, there was less objectivity in observing and evaluating both passing and failing performances. These findings are further reinforced by the fewer missed-event errors committed by raters who used the structured observation form.

A slightly different picture emerges, however, when the performance observation forms used in Experiment II are compared in terms of rater accuracy and reliability. In this experiment, performance on the structured observation form was similar to that on the semistructured form. The lack of difference found in the accuracy and reliability of judgments when using these forms may be due to the information-processing demands in the performance observation situation. That is, the oscilloscope operator tasks viewed in Experiment II were much more complex than the VOM operator tasks in that the procedural steps associated with the former are interchangeable and, to a certain extent, the control settings on the oscilloscope are interactive (i.e., an increase in the value of one control can be compensated for by a decrease in another).

The use of a detailed step-by-step observation form is more demanding in this situation because a rater's overall evaluation is based upon correct task procedure (i.e., the process by which measurements were taken) and not simply the task outcome (i.e., actual readings obtained). As a result, conflicting demands may be placed on a rater, who must continually scan the entire check-off sheet to ensure that a "yes" or "no" response is recorded for individual performance steps, while at the same time being careful to observe on-going task performance. The semistructured form, in contrast, minimized data-recording time for individual performance steps, thereby allowing more time for observation of the task itself.

These findings suggest that the level of task complexity should be considered when selecting the most appropriate performance evaluation procedure. Although the structured observation form resulted in superior judgment accuracy and reliability when a relatively simple task was evaluated in Experiment I, these advantages were offset in Experiment II when the complexity of the task and the number of interchangeable steps in the task increased. A trade-off may exist, therefore, between the information-processing demands associated with observing a highly complex task and the use of a highly detailed form to evaluate task performance. The exact level at which this trade-off occurs for other systems or equipments needs to be explored.

The Navy typically uses unstructured or semistructured forms to evaluate hands-on job performance. In fact, the Navy's Personnel Qualifications Standards (PQS) program is, to a great extent, based on the certification of job performance by means of unstructured rating forms. The current findings suggest that the use of more structured performance observation forms will produce more accurate, reliable, and objective measurements of hands-on job performance, at least for some tasks.

CONCLUSIONS

1. Within the range of rater skill tested in the current study, the anticipated results that raters who were more highly skilled in the operation of a particular type of electronic test equipment would be more accurate and consistent in evaluating the performance of others using the same equipment did not materialize. This means that having above average skill in a given task area does not automatically guarantee superior performance in terms of either rater accuracy or reliability.

2. The presence of at least some structure in a performance observation form produces more accurate and reliable evaluations of job task performance than when there is no structure.

3. An important consideration in the selection of an appropriate job task evaluation procedure is the level of complexity associated with the task to be evaluated. A trade-off may exist between the information-processing demands associated with observing a highly complex task and the use of a highly detailed form to evaluate job task performance. That is, the advantage in using a detailed step-by-step form to evaluate a task may be offset as either the complexity of the task or the number of interchangeable steps increases.

RECOMMENDATIONS

It is clear that the amount of structure provided in a performance observation form has an important effect on the accuracy and consistency of raters' judgments. Results from the current study have direct implications for both the design and development of new methods for evaluating the job performance proficiency of Navy personnel and the revision of existing methods.

The level of complexity associated with the tasks to be observed and evaluated is an important consideration in determining the amount of structure to be included in a performance observation form. The current findings suggest that, if the task to be evaluated is a simple one, the structured observation form should be used because it resulted in consistently higher rater performance (i.e., judgment accuracy and reliability) than the semistructured form. Further, as the complexity or number of interchangeable steps in a job task increases, the information-processing demands placed upon a rater by the performance observation form should be held to a minimum (as in the case of the semistructured form). It should be noted, however, that the current effort investigated job performance evaluation techniques based on only two equipment operator tasks. Therefore, further research should be conducted to determine the exact nature of the trade-off between task complexity and performance observation form specificity.

Use of the structured observation technique is not restricted to the evaluation of actual on-the-job performance. The use of this technique should be tested to determine its suitability for evaluating trainee proficiency in formal Navy training schools. This technique would not only ensure highly accurate and reliable judgments of trainee performance but would also provide trainees with diagnostic information about their specific strengths and weaknessess in performing equipment operating procedures. Such information is not directly retrievable from the semistructured and unstructured performance observation forms.

REFERENCES

Hays, W. L. Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Laabs, G. J., Panell, R.C., & Picker,ng. E. J. A personnel readiness training program: Maintenance of the missile test and readiness equipment (MTRE) MK 7 MOD 2) (NPRDC Tech. Rep. 77-19). San Diego: Navy Personnel Research and Development Center, March 1977. (AD-A037 546)

Laabs, G. J., Nugent, W. A., & Bearden, R. M. Validation of three versions of an oscilloscope operator test varying in fidelity. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, April 1981.

McCluskv, M. R., Trepagnier, J. C., Cleary, F. K., & Tripp, J. M. Evaluation of prototype job performance tests for the U. S. Army infantryman (Final Report CD-(C)-75-9). Alexandria, VA: Human Resources Research Organization for U. S. Army Research Institute for the Behavioral and Social Sciences, October 1975.

Pickering, E. J., & Anderson, A. V. Measurement of job-performance capabilities (NPRDC Tech. Rep. 77-6). San Diego: Navy Personnel Research and Development Center, December 1976. (AD-A033 992)

Scheffe, H. A. A method for judging all contrasts in the analysis of variance. Biometrika, 1953, 40, 87-104.

Shrout, P. E. & Fleiss, J. L. Interclass Correlations: Uses in assessing rater reliability. Psychological Bulletin, 1979, 86(2), 420-428.

Siegel, A. I. Retest-Reliability by a movie technique of test administrators' judgments of performance in process. Journal of Applied Psychology, 1954, 38, 390-392.

Snedecor, G. W. & Cochran, W. G. Statistical Methods (7th ed.). Ames, IW: Iowa State University Press, 1980.

APPENDIX A

EXAMPLES OF PERFORMANCE OBSERVATION FORMS USED IN EXPERIMENT I

Example of the Unstructured Observation Form

1. Voltage at Post 6 of the signal generator measured properly? □ Passed

□ Failed

What errors did you observe? ____________________________________

____

____

Example of the Semistructured Observation Form

PROBLEM #1 : VOM DC VOLTAGE MEASUREMENT

A. PRELIMINARY ADJUSTMENTS TOTAL_____

Zero Adjustment _____(1.0)

Range Selector Switch-highest DC _____(0.5)

Lead Connections _____(0.5)

B. CONTROL SETTING TOTAL_____

Sets Range Selector Switch to
most accurate DC range for
measurement (mid-scale) _____(1.0)

C. SOLUTION TOTAL_____

Meter reading accuracy (-14.4 VDC to -17.6 VDC)

Actual reading__________ _____(4.0)

D. SOLUTION TIME TOTAL_____

Allowed - 3 minutes

Actual__________ _____(1.0)

E. SAFETY _____(2.0) TOTAL_____

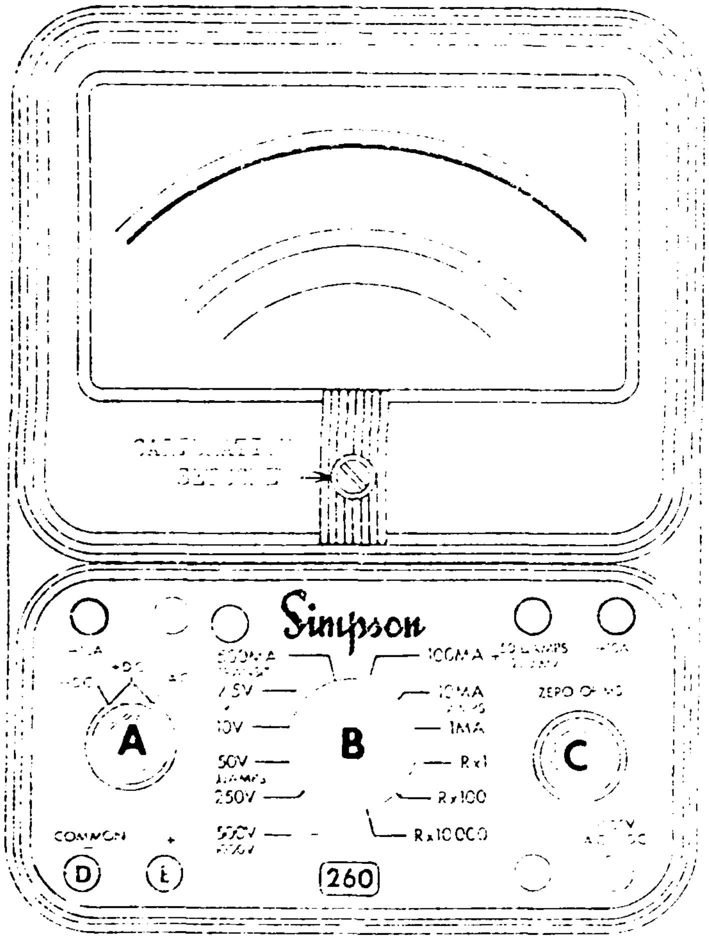
PROBLEM TOTAL_____

PASSED FAILED

□ □

(The total number of points necessary for passing is 7.5)

Example of the Structured Observation Form



Problem 1: DC VOLTAGE CHECK

Performed Correctl

METER CHECKOUT PROCEDURE

1. Was the meter set to zero using the calibration setscrew? YES NO

2. Was Switch B turned to the 500 V position? YES NO

3. Was the black test lead plugged into the jack labeled D (Common), and the red test lead plugged into the jack labeled E (+)? YES NO

NEGATIVE DC VOLTAGE MEASUREMENT

1. Was the black lead from the jack labeled D (Common) connected to a black post (ground) on the signal generator? YES NO

2. Was the red lead from the jack labeled E (+) connected to Post 6 of the signal generator? YES NO

3. Was the final position of Switch A set at -DC? YES NO

4. Was the final position of Switch B set at 50 V? YES NO

5. Was the voltage reading reported between -15 and -17 VDC? YES NO

PASSED FAILED

APPENDIX B

EXAMPLES OF PERFORMANCE OBSERVATION FORMS USED IN EXPERIMENT II

Example of the Unstructured Observation Form

1. Was the peak-to-peak amplitude of the signal at Test Point #1 measured properly?

□ Passed

□ Failed

What errors did you observe? ____________________

____

____

Example of the Semistructured Observation Form

PROBLEM 1 : AMPLITUDE MEASUREMENT

A. PRELIMINARY ADJUSTMENTS

Intensity/Focus
Input Coupling - AC/DC
Display - Channel A
Probe Connections Correct

MAXIMUM POINTS (4.0)
POINTS ASSIGNED: ________

B. CONTROL SETTINGS

Volts/Division - (.05 - .2 cm)
Time/Division - (1 - 20 μsec)
Trigger Level - Stable
Channel A Vernier - CAL

MAXIMUM POINTS (4.0)
POINTS ASSIGNED: ________

C. WAVEFORM ANALYSIS

Amplitude Allowed - (2.5 - 2.8 v)
Amplitude Reported ________

MAXIMUM POINTS (15.0)
POINTS ASSIGNED: ________

D. SAFETY

MAXIMUM POINTS (2.0)
POINTS ASSIGNED: ________

PROBLEM TOTAL ________

PASSED □ FAILED □

(The total number of points necessary for passing is 20.0)

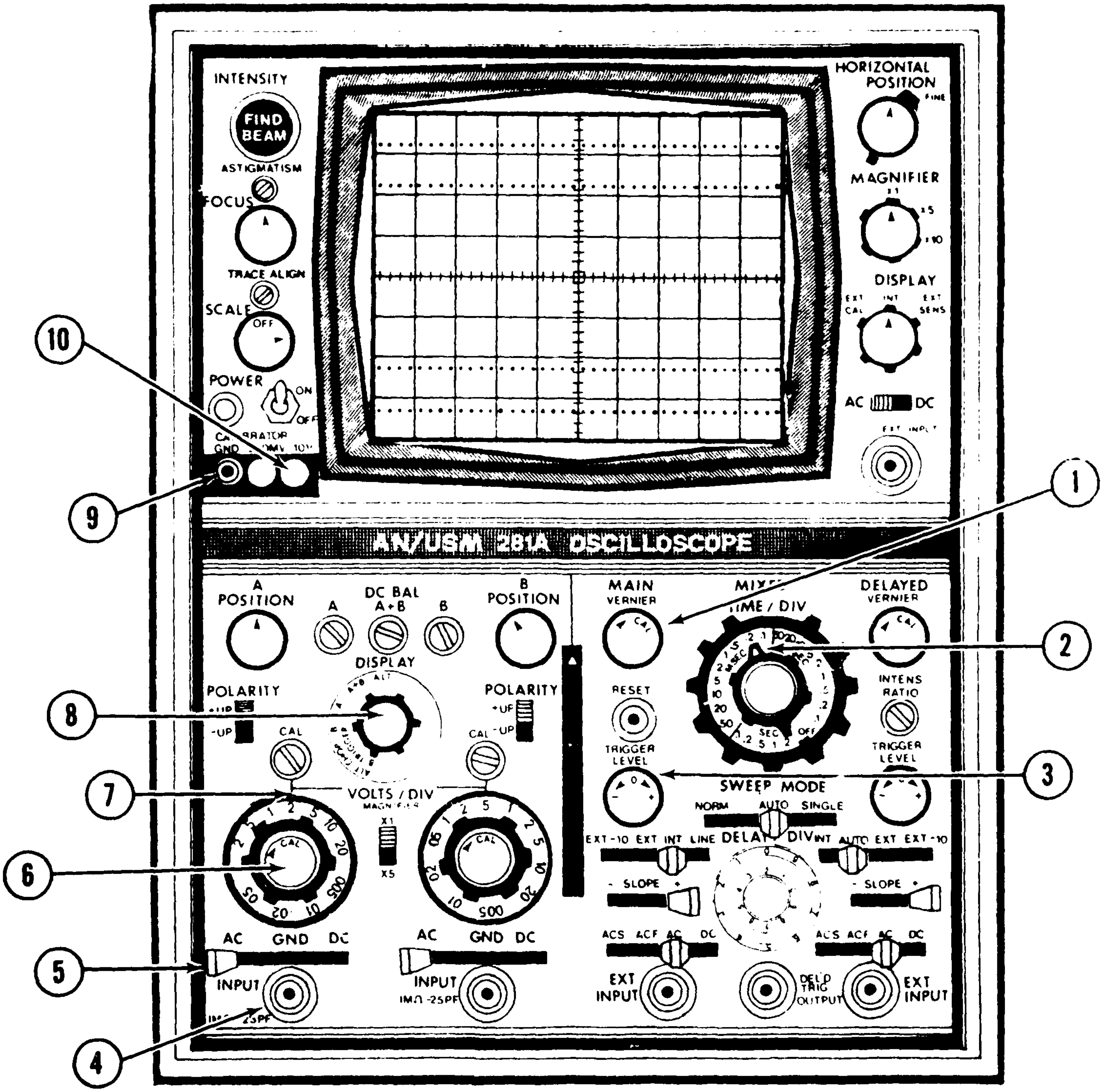
Example of the Structured Observation Form

PROBLEM 1 : AMPLITUDE MEASUREMENT

INITIAL SET-UP	PERFORMED CORRECTLY?	
1. Was control ⑧ set to the channel A position?	YES	NO
2. Was Switch ⑤ set to AC or DC?	YES	NO
3. Was the 10:1 probe connected to input jack ④, test point 1, and ground on the black box?	YES	NO

AMPLITUDE MEASUREMENT PROCEDURE		
1. Was the final position of Control ⑦ set between .05 and .2 centimeters (cm) deflection?	YES	NO
2. Was Control ⑥ set in the CAL position?	YES	NO
3. Was a stable waveform displayed (using Control ③ as necessary)?	YES	NO
4. Was the number of grid divisions reported between 1.3 and 5.2 centimeters (cm)?	YES	NO
5. Was the amplitude of the signal reported between 2.5 and 2.8 volts (v)?	YES	NO

PASSED □ FAILED □



INTENSITY
FIND BEAM
ASTIGMATISM
FOCUS
TRACE ALIGN
SCALE
OFF
POWER
ON
OFF
CALIBRATOR
GND
10MV
10V
HORIZONTAL POSITION
FINE
MAGNIFIER
X1
X5
X10
DISPLAY
EXT CAL
INT
EXT SENS
AC
DC
EXT INPUT
AN/USM 281A OSCILLOSCOPE
A POSITION
DC BAL
A
A+B
B
B POSITION
DISPLAY
A
A+B
ALT
POLARITY
+UP
-UP
CAL
VOLTS / DIV
MAGNIFIER
X1
X5
AC
GND
DC
INPUT
1MΩ-25PF
MAIN
VERNIER
CAL
MIXED
TIME / DIV
DELAYED
VERNIER
CAL
RESET
TRIGGER LEVEL
INTENS RATIO
TRIGGER LEVEL
SWEEP MODE
NORM
AUTO
SINGLE
EXT-10
EXT
INT
LINE
DELAY DIV
INT
AUTO
EXT
EXT-10
- SLOPE +
- SLOPE +
ACS
ACF
AC
DC
EXT INPUT
DEL'D TRIG OUTPUT
EXT INPUT
SEC
OFF
1
2
3
4
5
6
7
8
9
10

DISTRIBUTION LIST

Director of Manpower Analysis (ODASN (M))

Chief of Naval Operations (OP-01), (OP-11), (OP-12) (2), (OP-115) (2), (OP-140F2), (OP-987H)

Chief of Naval Material (NMAT 0722), (NMAT 08L)

Chief of Naval Research (Code 200), (Code 440) (3), (Code 442), (Code 448)

Chief of Information (OI-213)

Chief of Naval Education and Training (02), (003), (022), (N-2), (N-5), (N-9)

Chief of Naval Technical Training (016)

Commander Fleet Training Group, Pearl Harbor

Commander Naval Military Personnel Command (NMPC-013C)

Commander Training Command, U.S. Atlantic Fleet

Commander Training Command, U.S. Pacific Fleet

Commanding Officer, Fleet Anti-Submarine Warfare Training Center, Pacific

Commanding Officer, Fleet Training Center, San Diego

Commanding Officer, Naval Education and Training Program Development Center (Technical Library) (2)

Commanding Officer, Naval Education and Training Support Center, Pacific

Commanding Officer, Naval Health Sciences Education and Training Command

Commanding Officer, Naval Regional Medical Center, Portsmouth, VA (Attn: Medical Library)

Commanding Officer, Naval Technical Training Center, Corry Station (Code 101B)

Commanding Officer, Recruit Training Command (Academic Training Division)

Director, Naval Civilian Personnel Command

Director, Naval Education and Training Program Development Center Detachment, Great Lakes

Director, Naval Education and Training Program Development Center Detachment, Memphis

Director, Training Analysis and Evaluation Group (TAEG)

Officer in Charge, Central Test Site for Personnel and Training Evaluation Program

Superintendent, Naval Postgraduate School

Commander, Army Research Institute for the Behavioral and Social Sciences, Alexandria (PERI-ASL)

Chief, Army Research Institute Field Unit, Fort Harrison

Commander, Air Force Human Resources Laboratory, Brooks Air Force Base (Scientific and Technical Information Office)

Commander, Air Force Human Resources Laboratory, Lowry Air Force Base (Technical Training Branch)

Commander, Air Force Human Resources Laboratory, Williams Air Force Base (AFHRL/OT)

Commander, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base (AFHRL/LR)

Commanding Officer, U.S. Coast Guard Research and Development Center, Avery Point

Superintendent, U.S. Coast Guard Academy

Defense Technical Information Center (DDA) (12)

DATE

LMED

-8